Estimation of Obesity Levels Based On Eating Habits and Physical Condition Using Naive Bayes Classification

Aldair Perez[1]

[1] School of Computer Science and Engineering
[2] California State University San Bernardino, United States of America
[3] 008306045@coyote.csusb.edu

Abstract

Throughout the world, obesity is becoming an increasingly large health problem that affects the quality of life of millions of people. By looking at the factors that cause obesity, obesity and the health issues that come with it will be prevented. This study uses machine learning in order to predict obesity based upon a multitude of different factors from all parts of a person's life using data from the 2019 study of the same name. Within the dataset, it contains 16 features entailing health and lifestyle choices such as method of transportation, diet, amount of exercise, and usage of technology. The Naive Bayes classification model is used and inspected in order to pinpoint the likelihood of obesity based upon these features. The model is assessed based upon its precision, accuracy, F1-score, and recall. In the end the model demonstrated a precision of 98.89%, an accuracy of 80.14%, a recall of 70.29%, and an F1-score of 82.20%.

1. Introduction

Across the world, obesity is running rampant and is leading to health complications for millions. Conditions such as diabetes are often a grave consequence of obesity, with a study by the Journal of the American Heart Association linking obesity to 30-53 percent of new diabetes cases in the United States yearly. [3] Diabetes itself leads to long term complications and significantly harms the physical and mental well-being of those who have it. Common complications include kidney disease, eye damage, and neuropathy. [1] Furthermore, CPS1 and NHS data estimates that there are an estimated 374,239 obesity-related deaths per year within the United States. [4]

By identifying those at risk of becoming obese, people who could potentially suffer from the consequences of obesity could benefit from medical intervention in order to ensure their quality of life. Within this study, a Naive Bayes' Classifier will be implemented in order to predict obesity based on data from the UCI Machine Learning Repository 2019 survey. [2] Within this dataset are survey answers surrounding physical activity and eating habits as well as simple attributes such as height and weight.

The subsequent sections are outlined as follows: Section 2 discusses related work surrounding machine learning algorithms and their implementations on obesity prediction. Section 3 provides insight on the dataset and the steps taken to preprocess the data. Section 4 explains why Naive Bayes was chosen. Section 5 delves into the experiment's results and evaluates the model's effectiveness. Concluding, Section 6 discusses future research and summarizes the experiment.

2. Related Work

Due to machine learning's presence in the medical field, obesity prediction is a familiar topic that many researchers have tackled before.

One such example is Alberto Gutierrez-Gallego and company's cascade classifier model that combined multiple models in order to predict obesity. [5] Within their study, they combined gradient boosting, random forest, and logistic regression models in order to predict obesity based on factors such as age, sex, academic level, profession, smoking habits, and wine consumption. [5] Their results demonstrated that combining multiple models produced the highest level of accuracy compared to separate models.

In Bangladesh, Faria Ferdowsy and company applied nine different machine learning algorithms in order to predict obesity based on 1100 instances. [6] Within their study, they tested k-nearest neighbor, random forest, logistic regression, multilayer perceptron, support vector machine, naive Bayes, adaptive boosting, decision tree, and gradient boosting classifier based on data features such as height, weight, gender, diet, physical activity, and mental health. Their work showed that logistic regression was the most accurate of the models as it achieved the highest accuracy of 97.09%. The worst performing model they used was the gradient boosting algorithm with an accuracy of 64.08%.

Sri Astuti Thamrin and company assessed the ability of logistic regression, classification and regression trees, and naive Bayes on the RISKESDAS survey conducted by the Indonesian Ministry of Health. [7] The dataset focused on features such as marital status, age group, education, work category, diet, and physical activity. The study showed that logistic regression had the best performance of the three models.

In Brazil, Elias Rodriguez and company used decision trees, support vector machines, k-nearest neighbors, gaussian naive Bayes, multilayer perceptron, random forest, gradient boosting, and extreme gradient boosting in order to assess each model's performance on a survey collecting data on age, height, weight, physical activity, diet, and lifestyle habits. [8] In their study, they found that random forest was the most successful algorithm with a 78% accuracy, 79% precision, 78% recall, and 78% F1-score.

Zeyu Zheng and Karen Ruggiero worked to test four different machine learning models on predicting obesity within high school students. They tested binary logistic regression, improved decision tree, weighted k-nearest neighbor, and artificial neural network on nine health-related behaviors from the 2015 Youth Risk Behavior Surveillance System for the state of Tennessee. [9] They found that logistic regression performed the worst with an accuracy of 56% while IDT, KNN, and ANN did far better with accuracies of 80.23%, 88.82%, and 84.22% respectively.

Overall, we can find throughout these studies that there are many different scenarios in which machine learning can be applied to predict harmful conditions such as obesity. The studies also show the sheer variety of different models and features that can be used in order to ensure the best results for these machine learning algorithms.

3. Datasets and Preprocessing

This study uses data from the Estimation of Obesity Levels Based on Eating Habits and Physical Condition 2019 study conducted by Fabio Mendoza Palechor and Alexis De la Hoz Manotas. [2] The original dataset consists of 2,111 survey responses with 16 features and 1 target. All features were directly answered by the respondents and they were classified into seven categories: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III.

Before preprocessing, the dataset is slightly imbalanced, as "Normal Weight" and "Overweight Level I" are overrepresented compared to the other categories. During preprocessing, the "method of transportation" feature was omitted due to it not being ordinal and not smoothly fitting in alongside the other features within the classifier. The data was fortunately already well processed, all that was necessary for the features was integer encoding in order to sort the ordinal features in numerical format for the classifier. Following this, the data was normalized using a Z-score standardization. Finally, the classes were simplified into Obese or Normal in order to simplify the classifier and just test for Obesity identification. Following preprocessing, the dataset became more balanced as about 54 percent were considered not obese and 46 percent were considered obese. The preprocessed dataset prior to normalization is visible on Table 1.

4. Methodologies

The algorithm chosen to be implemented was the Naive Bayes Classifier. This algorithm was chosen due to its ease of use and low computational cost.

4.1. Naive Bayes Classifier

The Naive Bayes Classifier is a supervised learning algorithm based on Bayes' theorem that functions on a "naive" expectation of conditional independence. [11] The term naive is used due to the fact that conditional independence is almost never the case in practice. Within this implementation, the Gaussian Naive Bayes variant is used in order to accommodate the continuous features within the dataset. Bayes' theorem is used to calculate posterior probability by combining the prior probability with the posterior (the probability of the class in general and the probability that certain features belong to a class). After computing these probabilities, the class with the highest total probability for each instance is selected and stored. The main function of the implementation functions as follows: for each class, compute log prior, compute log of Gaussian probability density function for each future, sum each log of Gaussian PDF to get value of log posterior, add log prior to log posterior, and choose the class with the highest probability after these steps to be stored in posteriors. We take log to avoid underflow, improve computation, and simplify the process by turning steps that would be multiplicative into additive steps.

5. Experimental Results and Discussion

To begin, the dataset was shuffled using a seed with Numpys in order to ensure replicability while still making sure the data grabs a fair amount from each class. The data was then split between training and testing. The split for this dataset was 80% training (1688 instances) and 20% testing (423 instances). The manually coded Naive Bayes Classifier fits the model to said training data and then uses its predict function to predict the test data.

Table 1. Feature Description

Feature	Type	Values	Description
Age	Continuous	18–60	Age of the individual in years
Gender	Categorical	0-1	Biological sex of the individual (0 = Female, 1 = Male)
Height	Continuous	1.4–2.0	Height in meters
Weight	Continuous	40–130	Weight in kilograms
FHOW	Binary	0-1	Whether family has a history with being overweight (0 = no, 1 = yes)
FAVC	Binary	0-1	Frequent consumption of high-calorie food (0 = no, 1 = yes)
FCVC	Ordinal	1–3	Daily intake of vegetables (1 = low, 3 = high)
NCP	Ordinal	1–4	Number of main meals per day
CAEC	Categorical	0-3	Consumption of food between meals (0 = none, 3 = high)
SMOKE	Binary	0-1	Whether or not the individual smokes (0 = no, 1 = yes)
CH2O	Ordinal	1–3	Daily water intake (1 = low, 3 = high)
SCC	Binary	0-1	Whether or not individual monitors calorie consumption (0 = no, 1 = yes)
FAF	Ordinal	0–3	Physical activity frequency (0 = none, 3 = high)
TUE	Ordinal	0–2	Time spent on technology per day in hours
CALC	Categorical	0-3	Alcohol consumption frequency (0 = none, 3 = high)
Obese	Binary (Target)	0-1	Obesity classification label (target variable) (0 = no, 1 = yes)

5.1. Naive Bayes Evaluation

Upon training and being evaluated, the Naive Bayes Classifier achieved an accuracy of 80.14 percent. The model achieved a precision of 98.89 percent, which means that the model did an excellent job of ensuring it only labeled obese people as obese (few false positives). Furthermore, the model achieved a recall of 70.29 percent, which means that the model was poor in identifying all cases of obesity (many false negatives). Finally, the model achieved an F1-score of 82.20 percent.

True Class	Predicted Class	
	0	1
0	145	2
1	82	194

Table 2. Confusion matrix of the Naive Bayes classifier.

Accuracy	Precision	Recall	F1
80.14%	98.89%	70.29%	82.20%

Table 3. The classification report of the Naive Bayes Classifier.

6. Conclusion

Within this study, Naive Bayes was implemented in order to predict obesity using the UCI Machine Learning dataset of the same name. The model achieved a fairly high accuracy of 80%. However, the model suffered with a lower recall of 70% which meant that there were too many positive cases of obesity slipping through the model's cracks.

Due to the self-reported nature of the questions asked, there is a high chance of embellishments and inaccuracies within the responses which could lead to decreased performance for machine learning models such as the Naive Bayes classifier tested. As a result of similarities in classes such as Overweight level II and Obesity level I, the model's binary labeling of very overweight people as not obese could've influenced it by causing it to draw similarities between people the next class up and failed to identify them properly. In the future, improving the model to identify each separate weight class could improve its results.

References

[1] World Health Organization, “Diabetes,” Nov. 2024.

[2] Fabio Mendoza Palechor, Alexis De la Hoz Manotas, “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico,” 2019.

[3] Journal of the American Heart Association, “Obesity contributes to up to half of new diabetes cases annually in the United States,” Feb. 2021.

[4] Eugenia Thoenen, “Obesity: Facts, Figures, Guidelines,” *Section One - OBESITY AND MORTALITY*, Dec. 2002.

[5] Alberto Gutierrez-Gallego et al., “Combination of Machine Learning Techniques to Predict Overweight/Obesity in Adults,” July, 2024.

[6] Faria Ferdowsy et al., “A machine learning approach for obesity risk prediction,” May, 2021.

[7] Sri Astuti Thamrin et al., “Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research,” June, 2021.

[8] Elias Rodríguez et al., “Machine learning techniques to predict overweight or obesity ,” Nov. 2021.

[9] Zeyu Zheng, Karen Ruggiero, “Using machine learning to predict obesity in high school students ,” Nov. 2017.

[10] Christoper D. Manning, Prabhakar Raghavan, Hinrich Schutze “Introduction to Information Retrieval ,” *Section 13 - Text classification and Naive Bayes*, July, 2008.

[11] Kilian Weinberger, “Estimating Probabilities from Data: Naive Bayes ,” July, 2018.

Address for correspondence:

Aldair Perez
https://github.com/tevezr7
008306045@coyote.csusb.edu